Statistics

|From Samples to Significance (الاستدلال)

Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

Today's Focus

- Quick Refresher
- Descriptive vs. Inferential Statistics
- Statistic vs. Parameter
- Randomness and Variability
- Margin of Error
- What Is Statistical Significance?
- Examples

🔁 Recap of Key Concepts

- A **subject** (also called a unit or case) is the individual entity from which data is collected in a statistical study.
- It could be a person, object, event, or organization, depending on the context of the study.

🔁 Recap of Key Concepts

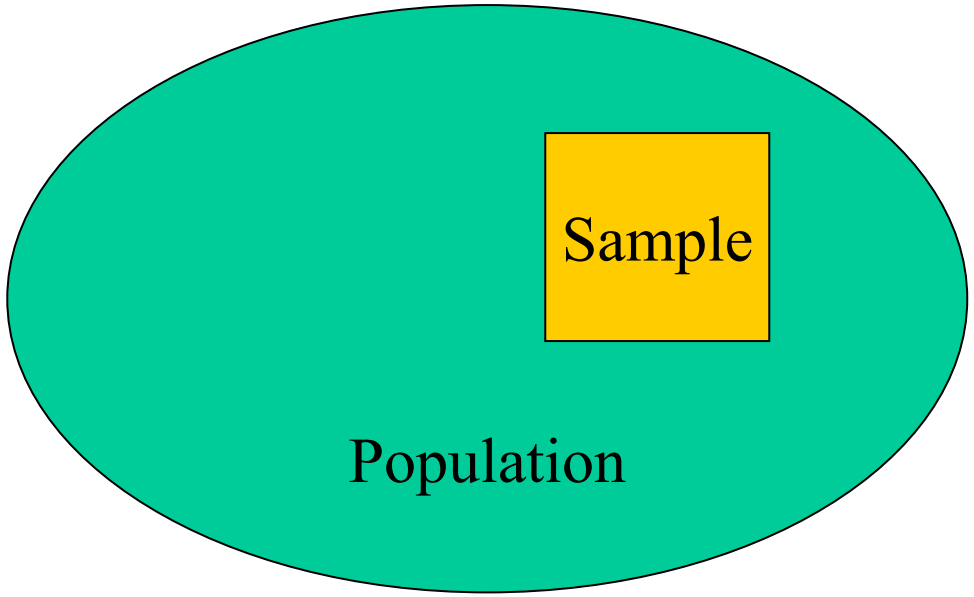
- **Subject**

Field / Industry	Study Focus	Subject
🏥 Healthcare	Effect of a new drug	Each **patient** in the clinical trial
🏫 Education	Student test performance and study hours	Each **student** in the class
🏢 Human Resources	Employee satisfaction survey	Each **employee** surveyed
🏠 Real Estate	Predicting house prices	Each **house/property** listed
🛍️ Retail	Customer shopping behavior	Each **customer** making a purchase
🚗 Automotive	Fuel efficiency of different car models	Each **car** tested
🌍 Environmental Science	Air quality monitoring across cities	Each **city** measured
📈 Finance	Stock performance analysis	Each **stock** or **company**

Recap of Key Concepts

- **Population**: The entire group you want to study or draw conclusions about.

- **Sample**: A subset of the population selected for actual data collection and analysis.



🔁 Recap of Key Concepts

Field	Population	Sample
🏥 Healthcare	All patients with diabetes in a country	500 diabetic patients from selected hospitals
🏫 Education	All high school students in a city	200 students from 5 randomly chosen schools
🛒 Market Research	All customers of an online store	1,000 customers who responded to a satisfaction survey
🏠 Real Estate	All houses sold in a region in 2024	A dataset of 150 homes sold in March 2024
📊 Finance	All companies listed on the stock market	50 companies selected from the technology sector
🌍 Environmental	All lakes in a country	30 lakes tested for pollution levels

📊 Descriptive vs. Inferential Statistics

- **Descriptive Statistics** refers to methods for **summarizing the collected** data. Summaries consist of graphs and numbers such as averages and percentages.
- **Inferential statistics** refers to methods of **making decisions or predictions** about a population based on data obtained from a sample of that population.

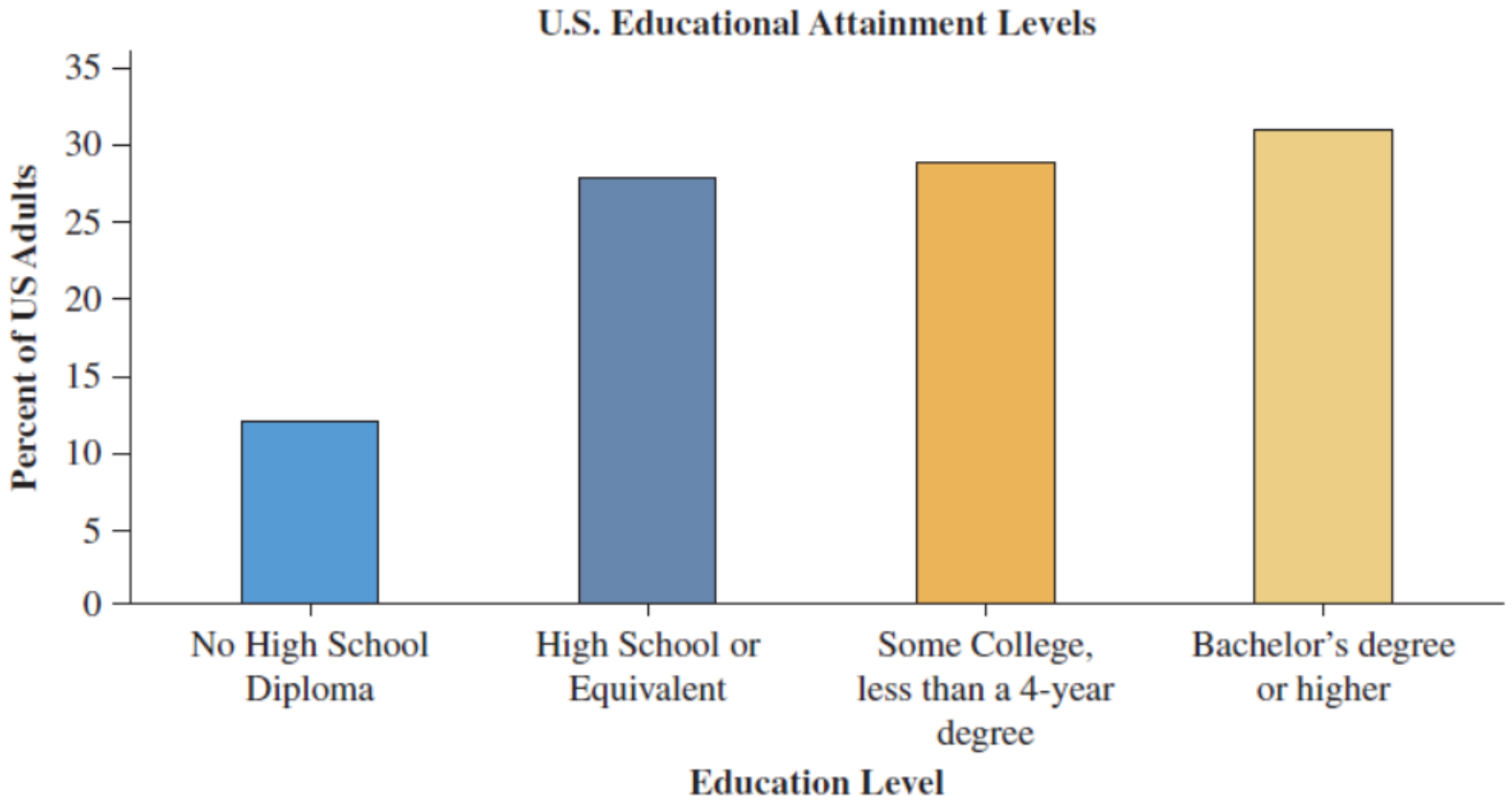
🧠 **In Simple Terms**:

- Descriptive statistics = What the data shows 📋
- Inferential statistics = What we can learn or predict from the data 🔍

📊 Descriptive vs. Inferential Statistics

Aspect	Descriptive Statistics	Inferential Statistics
Definition	Summarizes and presents data collected from a group	Draws conclusions or predictions about a population based on a sample
Goal	Describe what the data shows	Make generalizations beyond the observed data
Data Scope	Focused on the sample or dataset in hand	Applies sample results to the larger population
Methods Used	Tables, charts, averages, variability (mean, median, SD)	Hypothesis testing, confidence intervals, regression analysis
Example Use	Showing the average exam score of students in one class	Estimating the average score of all students in the school
Question Answered	"What does the data tell us?"	"What can we infer or predict from the data?"

Descriptive Statistics Example



▲ **Figure 1.1** Educational Attainment, Based on a Sample of 78,000 Households in the 2013 Current Population Survey. (*Source:* Data from United States Census Bureau.)

Inferential Statistics Example

- Polling opinions on Handgun Control - control over the sales of handguns.
- study results from a recent poll of 834 Florida residents
  - In that poll, 54.0% of the sampled subjects said they favored controls over the sales of handguns.
  - We are 95% confident that the percentage of all adult Floridians favoring control over sales of handguns falls between 50.6% and 57.4%.

📊 Statistic vs. Parameter

- **Statistic**: A numerical value that describes a sample. It is calculated from collected data.
  - 54.0% of the sampled subjects said they favored controls over the sales of handguns.
- **Parameter**: A numerical value that describes a population. It is usually unknown and estimated using a statistic.
  - Percentage of the population of all Florida residents favoring handgun control

📊 Statistic vs. Parameter

Aspect	Statistic	Parameter
Definition	A value calculated from a **sample**	A value that describes the **entire population**
Data Source	Sample (subset of the population)	Population (entire group)
Known/Unknown	Known (we compute it directly)	Often unknown (we estimate it)
Use	To estimate the parameter	Represents the true value we aim to know
Symbols	Latin letters (e.g., $\bar{x}$, s)	Greek letters (e.g., μ, σ)
Example	The average income of 100 surveyed people = $\bar{x}$	The actual average income of all citizens = μ

🎯 Randomness and Variability

- **Randomness** means that outcomes are not predictable but follow a known probability pattern.
- Randomness is essential for:
  - Designing fair experiments
  - Ensuring unbiased samples
  - Making valid generalizations (inferences) about a population

🎯 Randomness and Variability

✅ Random sampling ensures that every subject in the population has an equal chance of being selected.

📌 Writing names on slips of paper, placing them in a hat, and drawing randomly — this is called a **Simple Random Sample** (SRS).

🎯 Randomness and Variability

📈 **Variability** is the natural differences observed between individuals or samples.

- Not all individuals are the same
- Repeated samples will produce slightly different results (**vary from sample to sample**)

✅ Understanding variability helps us measure **how precise or reliable** our results are.

That's why tools like **margin of error** and **standard deviation** are important.

📉 Margin of Error

- with **random sampling**, the amount of **variability** from sample to sample is actually **quite predictable**
- The **margin of error** is a measure of the expected variability from one random sample to another.
  - It tells us how much a sample estimate might differ from the true population value due to random chance.
  - Approximate margin of error $= \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}} \times \mathbf{100}\ \%$

🔢 Approximate Formula

Margin of Error (%) $\approx \frac{1}{\sqrt{n}} \times 100\ \%$

- Where n = sample size

👉 A larger sample = smaller margin of error

🧠 Example:

A poll shows that 52% of respondents support a new policy, with a margin of error ±3%.

✅ Interpretation: It is very likely that the true population percentage is between 49% and 55%.

📊 What Is Statistical Significance?

- **Statistical significance** means that the difference observed in a study is so large that it is **unlikely** to have occurred **by random chance alone**.

✅ When the probability of such a difference happening by chance is very low (typically less than 5%), we call the result statistically significant.

🧪 Example: Chemotherapy Trial

- A **new chemotherapy** drug is tested against a standard treatment:
  - If patients receiving the new treatment show a **much better outcome** (e.g., higher survival or lower recurrence),
  - And this difference is **unlikely to result** from random variation,

➡️ Then we can conclude that the new chemotherapy is effective, based on **statistical significance**.

📊 Example 01 - New public health policy

🎯 Scenario:

A research team surveys 400 people to find out what percentage support a new public health policy.

- Result from sample: 62% support the policy

- Sample size (n): 400
- Approximate Margin of Error:
- MOE (%) $\approx \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}} \times \mathbf{100}\ \% = \frac{\mathbf{1}}{\sqrt{400}} \times \mathbf{100}\ \% = 5\%$

📊 Example 01 - New public health policy

🎯 Scenario:

A research team surveys 400 people to find out what percentage support a new public health policy.

- Result from sample: 62% support the policy

✅ Interpretation:

We say the percentage of support in the population is approximately between 57% and 67% (i.e., 62% ± 5%).

🧠 Meaning:

- It’s very likely (typically 95% confidence) that if we surveyed the entire population, the true percentage would fall within this range.

📊 Example 02 - Customer Satisfaction

🎯 Scenario:

A company surveys a random sample of 1,000 customers to assess satisfaction with a newly launched product.

- Result from sample: 78% of customers are satisfied

- Sample size (n): 1,000
- Approximate Margin of Error:
- MOE (%) $\approx \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}} \times \mathbf{100}\ \% = \frac{\mathbf{1}}{\sqrt{1000}} \times \mathbf{100}\ \% = 3.16\%$

📊 Example 02 - Customer Satisfaction

🎯 Scenario:

A company surveys a random sample of 1,000 customers to assess satisfaction with a newly launched product.

- Result from sample: 78% of customers are satisfied

$$\text{MOE (\%)} \approx \frac{1}{\sqrt{n}} \times 100\ \% = \frac{1}{\sqrt{1000}} \times 100\ \% = 3.16\%$$

✅ Interpretation:

- The true satisfaction rate in the entire customer base is estimated to be between 74.84% and 81.16%(i.e., 78% ± 3.16%)